Two algorithms for fitting constrained marginal models

R. J. Evans, Statistical Laboratory, University of Cambridge, UK
A. Forcina, Dipartimento di Economia, Finanza e Statistica, University of Perugia, Italy

May 3, 2012

Abstract

There are two main algorithms which have been considered for fitting constrained marginal models to discrete data; these are studied in detail and their properties clarified. The two procedures are shown to be equivalent, in the sense that the updates they produce are identical, each method being advantageous in different circumstances. An extension is provided to one of the algorithms for modelling the effect of exogenous individual-level covariates, and an application of the method to likelihood-based estimation under L_1-penalties is also considered.

Keywords: categorical data, L_1-penalty, marginal log-linear model, maximum likelihood, non-linear constraint.

1 Introduction

The application of marginal constraints to multi-way contingency tables has been much investigated in the last 20 years; see, for example, McCullagh and Nelder (1989), Liang et al. (1992), Glonek and McCullagh (1995), Agresti (2002), Bergsma et al. (2009). Bergsma and Rudas (2002) introduced marginal log-linear parameters (MLLPs), which generalize other discrete parameterizations including ordinary log-linear parameters and Glonek and McCullagh's multivariate logistic parameters. The flexibility of this family of parameterizations enables their application to many popular classes of conditional independence models, and especially to graphical models (Forcina et al., 2010, Rudas et al., 2010, Evans and Richardson, 2011). Bergsma and Rudas (2002) show that, under certain conditions, models defined by linear constraints on MLLPs are curved exponential families. However, naïve algorithms for maximum likelihood estimation with MLLPs face several challenges: in general, there are no closed form equations for computing raw probabilities from MLLPs, so direct evaluation of the log-likelihood can be time consuming. In addition, MLLPs are not necessarily variation independent and, as noted by Bartolucci et al. (2007), ordinary Newton-Raphson or Fisher scoring methods may become stuck by producing updated estimates which are incompatible.

Lang (1996) and Bergsma (1997), amongst others, have tried to adapt a general algorithm introduced by Aitchison and Silvey (1958) for constrained maximum likelihood estimation to the context of marginal models. In this paper we provide an explicit formulation of Aitchison and Silvey's algorithm, and show that an alternative method due to Colombi and Forcina (2001) is equivalent; we term this second approach the *regression algorithm*. The regression algorithm may be preferable if the number of constraints is large, particularly in the presence of individual-level covariates, in which case Aitchison and Silvey's approach is often infeasible. A variation of these algorithms, which can be used to fit marginal log-linear models under L_1-penalties, is also given.

In Section 2 we review marginal log-linear models and their basic properties. In Section 3 we formulate the two algorithms, show that they are equivalent and discuss their properties. In Section 4 we describe applications to individual-level covariates, and finally Section 5 considers similar methods for L_1-constrained estimation.

2 Notations and preliminary results

Let X_j, $j = 1, \dots, d$ be categorical random variables taking values in $\{1, \dots, c_j\}$. The joint distribution of $X_1, \dots, X_d$ is determined by the vector of joint probabilities $\boldsymbol{\pi}$, of dimension $t = \prod_1^d c_j$, whose entries correspond to cell probabilities, and are assumed to be strictly positive; we take the entries of $\boldsymbol{\pi}$ to be in lexographic order. Further, let $\boldsymbol{y}$ denote the vector of cell frequencies with entries arranged in the same order as $\boldsymbol{\pi}$. We write the multinomial log-likelihood in terms of the canonical parameters as

$$l(\boldsymbol{\theta}) = \boldsymbol{y}'\boldsymbol{G}\boldsymbol{\theta} - n \log[\boldsymbol{1}_t' \exp(\boldsymbol{G}\boldsymbol{\theta})]$$

(see, for example, Bartolucci et al., 2007, p. 699); here n is the sample size, $\boldsymbol{1}_t$ a vector of length t whose entries are all 1, and $\boldsymbol{G}$ a $t \times (t-1)$ full rank design matrix which determines the log-linear parameterization. The mapping between the canonical parameters and the joint probabilities may be expressed as

$$\log(\boldsymbol{\pi}) = \boldsymbol{G}\boldsymbol{\theta} - \boldsymbol{1}_t \log[\boldsymbol{1}_t' \exp(\boldsymbol{G}\boldsymbol{\theta})] \quad \Leftrightarrow \quad \boldsymbol{\theta} = \boldsymbol{L} \log(\boldsymbol{\pi}),$$

where $\boldsymbol{L}$ is a $(t-1) \times t$ matrix of row contrasts and $\boldsymbol{L}\boldsymbol{G} = \boldsymbol{I}_{t-1}$.

Expressions for the score vector, $\boldsymbol{s}$, and the expected information matrix, $\boldsymbol{F}$, with respect to $\boldsymbol{\theta}$ take the form

$$\boldsymbol{s} = \boldsymbol{G}'(\boldsymbol{y} - n\boldsymbol{\pi}) \qquad \text{and} \qquad \boldsymbol{F} = n\boldsymbol{G}'\boldsymbol{\Omega}\boldsymbol{G};$$

here $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$.

2.1 Marginal log-linear parameters

Marginal log-linear parameters (MLLPs) enable the simultaneous modelling of several marginal distributions (see, for example, Bergsma et al., 2009, Chapters 2 and 4) and the specification of suitable conditional independencies within marginal distributions of interest (see Evans and Richardson, 2011). In the following let $\boldsymbol{\eta}$ denote an arbitrary vector of MLLPs; it is well known that this can be written as

$$\boldsymbol{\eta} = \boldsymbol{C} \log(\boldsymbol{M}\boldsymbol{\pi}),$$

where $\boldsymbol{C}$ is a suitable matrix of row contrasts, and $\boldsymbol{M}$ a matrix of 0's and 1's producing the appropriate margins (see, for example, Bergsma et al., 2009, Section 2.3.4).

Bergsma and Rudas (2002) have shown that if a vector of MLLPs $\boldsymbol{\eta}$ is *complete* and *hierarchical*, two properties defined below, models determined by linear restrictions on $\boldsymbol{\eta}$ are curved exponential families, and thus smooth. Like ordinary log-linear parameters, MLLPs may be grouped into interaction terms involving a particular subset of variables; each interaction term must be defined within a margin of which it is a subset.

Definition 1. *A vector of MLLPs $\boldsymbol{\eta}$ is called* complete *if every possible interaction is defined in precisely one margin.*

Definition 2. *A vector of MLLPs $\boldsymbol{\eta}$ is called* hierarchical *if there is a non-decreasing ordering of the margins of interest $M_1, \dots, M_s$ such that, for each $j = 1, \dots s$, no interaction term which is a subset of M_j is defined within a later margin.*

3 Two algorithms for fitting marginal log-linear models

Here we describe the two main algorithms used for fitting models of the kind described above.

3.1 An adaptation of Aitchison and Silvey's algorithm

Aitchison and Silvey (1958) study maximum likelihood estimation under non-linear constraints in a very general context, and show that, under certain conditions, the maximum likelihood estimates exist and are asymptotically normal; they also outline an algorithm for computing those estimates. Suppose we wish to maximize $l(\boldsymbol{\theta})$ subject to $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{0}$, a set of r non-linear constraints, under the assumption that the second derivative of $\boldsymbol{h}(\boldsymbol{\theta})$ exists and is bounded. Aitchison and Silvey propose to maximize the function $l(\boldsymbol{\theta}) + \boldsymbol{h}(\boldsymbol{\theta})'\boldsymbol{\lambda}$, where $\boldsymbol{\lambda}$ is a vector of Lagrange multipliers; this leads to the system of equations

$$\begin{aligned} \boldsymbol{s}(\hat{\boldsymbol{\theta}}) + \boldsymbol{H}(\hat{\boldsymbol{\theta}})\hat{\boldsymbol{\lambda}} &= \boldsymbol{0} \\ \boldsymbol{h}(\hat{\boldsymbol{\theta}}) &= \boldsymbol{0}, \end{aligned} \tag{1}$$

where $\hat{\boldsymbol{\theta}}$ is the ML estimate and $\boldsymbol{H}$ the derivative of $\boldsymbol{h}'$ with respect to $\boldsymbol{\theta}$. Since these are non-linear equations, they suggest an iterative algorithm which proceeds as follows: suppose that at the current iteration we have $\boldsymbol{\theta}_0$, a value reasonably close to $\hat{\boldsymbol{\theta}}$. Replace $\boldsymbol{s}$ and $\boldsymbol{h}$ with first order approximations around $\boldsymbol{\theta}_0$; in addition replace $\boldsymbol{H}(\hat{\boldsymbol{\theta}})$ with $\boldsymbol{H}(\boldsymbol{\theta}_0)$ and the second derivative of the log-likelihood with $-\boldsymbol{F}$, minus the expected information matrix. The resulting equations, after rearrangement, may be written in matrix form as

$$\begin{pmatrix} \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \hat{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{F}_0 & -\boldsymbol{H}_0 \\ -\boldsymbol{H}_0' & \boldsymbol{0} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{s}_0 \\ \boldsymbol{h}_0 \end{pmatrix},$$

where $\boldsymbol{s}_0$, $\boldsymbol{F}_0$, $\boldsymbol{H}_0$ denote the corresponding quantities evaluated at $\boldsymbol{\theta}_0$. To compute a solution, Aitchison and Silvey (1958) exploit the structure of the partitioned matrix, while Bergsma (1997) solves explicitly for $\hat{\boldsymbol{\theta}}$ by substitution; in both cases, if we are uninterested in the Lagrange multipliers, we get the updating equation

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \boldsymbol{F}_0^{-1}\boldsymbol{s}_0 - \boldsymbol{F}_0^{-1}\boldsymbol{H}_0(\boldsymbol{H}_0'\boldsymbol{F}_0^{-1}\boldsymbol{H}_0)^{-1}(\boldsymbol{H}_0'\boldsymbol{F}_0^{-1}\boldsymbol{s}_0 + \boldsymbol{h}_0). \tag{2}$$

As noted by Bergsma (1997), the algorithm does not always converge unless some sort of step length adjustment is introduced.

Linearly constrained marginal models are defined by $\boldsymbol{K}'\boldsymbol{\eta} = \boldsymbol{0}$, where $\boldsymbol{K}$ is a matrix of full column rank $r \leq t - 1$. The multinomial likelihood is a regular exponential family, so these models may be fitted using the smooth constraint $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{K}'\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{0}$, which implies that

$$\boldsymbol{H}' = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}'} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\eta}'}\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}'} = \boldsymbol{K}'\boldsymbol{R}^{-1},$$

where

$$\boldsymbol{R} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\eta}'} = \left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\theta}'}\right)^{-1} = [\boldsymbol{C}\,\mathrm{diag}(\boldsymbol{M}\boldsymbol{\pi})^{-1}\boldsymbol{M}\,\mathrm{diag}(\boldsymbol{\pi})\boldsymbol{G}]^{-1}.$$

Remark 1. *In the equation above, the inverse derivative is used because $\boldsymbol{\theta}$ cannot be written as a closed form function of $\boldsymbol{\eta}$. In this expression we have replaced $\boldsymbol{\Omega}$ with* $\mathrm{diag}(\boldsymbol{\pi})$ *by* *exploiting the fact that $\boldsymbol{\eta}$ is a homogeneous function of $\boldsymbol{\pi}$ (see Bergsma et al., 2009, Section 2.3.4). If the constrained model were not smooth then at singular points the Jacobian matrix $\boldsymbol{R}$ would not be invertible, implying that $\boldsymbol{H}$ is not of full rank and thus violating a crucial assumption in Aitchison and Silvey (1958). It has been shown (Bergsma and Rudas, 2002, Theorem 3) that completeness is a necessary condition for smoothness.*

3.2 A regression algorithm

By noting that the Aitchison-Silvey algorithm is essentially based on a quadratic approximation of $l(\boldsymbol{\theta})$ with a linear approximation of the constraints, Colombi and Forcina (2001) designed an algorithm which they believed to be equivalent to the original, though no formal argument was provided; this equivalence is proven in Proposition 1 below. Recall that, by elementary linear algebra, there exists a $(t-1)\times(t-r-1)$ design matrix $\boldsymbol{X}$ of full column rank such that $\boldsymbol{K}'\boldsymbol{X}=\boldsymbol{0}$, from which it follows that $\boldsymbol{\eta}=\boldsymbol{X}\boldsymbol{\beta}$ for a vector of $t-r-1$ unknown parameters $\boldsymbol{\beta}$. Let $\bar{\boldsymbol{s}}=\boldsymbol{R}'\boldsymbol{s}$ and $\bar{\boldsymbol{F}}=\boldsymbol{R}'\boldsymbol{F}\boldsymbol{R}$ respectively denote the score and information relative to $\boldsymbol{\eta}$; then the *regression algorithm* consists of alternating the following steps:

1. update the estimate of $\boldsymbol{\beta}$ by
$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0=(\boldsymbol{X}'\bar{\boldsymbol{F}}_0\boldsymbol{X})^{-1}\boldsymbol{X}'(\bar{\boldsymbol{F}}_0\boldsymbol{\gamma}_0+\bar{\boldsymbol{s}}_0), \tag{3}$$
where $\boldsymbol{\gamma}_0=\boldsymbol{\eta}_0-\boldsymbol{X}\boldsymbol{\beta}_0$;
2. update $\boldsymbol{\theta}$ by
$$\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0=\boldsymbol{R}_0[\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)-\boldsymbol{\gamma}_0]. \tag{4}$$

Proposition 1. *The updating equation in (2) is equivalent to the combined steps given in (3) and (4).*

A proof of this result is given in A.

Remark 2. *From the form of the updating equations (2), (3) and (4) it is clear that Proposition 1 remains true if identical step length adjustments are applied to the $\boldsymbol{\theta}$ updates. This does not hold, however, if adjustments are applied to the $\boldsymbol{\beta}$ updates of the regression algorithm.*

3.2.1 Derivation of the regression algorithm

In a neighbourhood of $\boldsymbol{\theta}_0$, approximate $l(\boldsymbol{\theta})$ by a quadratic function Q having the same information matrix and the same score vector as $l(\boldsymbol{\theta})$,

$$l(\boldsymbol{\theta})\cong Q(\boldsymbol{\theta})=-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{t}_0)'\boldsymbol{F}_0(\boldsymbol{\theta}-\boldsymbol{t}_0),\qquad\text{where}\qquad \boldsymbol{t}_0=\boldsymbol{\theta}_0+\boldsymbol{F}_0^{-1}\boldsymbol{s}_0.$$

Now compute a linear approximation of $\boldsymbol{\theta}$ with respect to $\boldsymbol{\beta}$ in a neighbourhood of $\boldsymbol{\theta}_0$,

$$\boldsymbol{\theta}-\boldsymbol{\theta}_0\cong\boldsymbol{R}_0(\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\eta}_0); \tag{5}$$

substituting into the expression for Q we obtain a quadratic function in $\boldsymbol{\beta}$. By adding and subtracting $\boldsymbol{R}_0\boldsymbol{X}\boldsymbol{\beta}_0$ and setting $\boldsymbol{\delta}=\boldsymbol{\beta}-\boldsymbol{\beta}_0$, we have

$$Q(\boldsymbol{\beta})=-\frac{1}{2}[\boldsymbol{R}_0\boldsymbol{X}\boldsymbol{\delta}-\boldsymbol{R}_0\boldsymbol{\gamma}_0-\boldsymbol{F}_0^{-1}\boldsymbol{s}_0]'\boldsymbol{F}_0[\boldsymbol{R}_0\boldsymbol{X}\boldsymbol{\delta}-\boldsymbol{R}_0\boldsymbol{\gamma}_0-\boldsymbol{F}_0^{-1}\boldsymbol{s}_0].$$

A weighted least square solution of this local maximization problem gives (3); substitution into (5) gives (4).

Remark 3. *The choice of $\boldsymbol{X}$ is somewhat arbitrary because the design matrix $\boldsymbol{XA}$, where $\boldsymbol{A}$ is any non-singular matrix, implements the same set of constraints as $\boldsymbol{X}$. In many cases an obvious choice for $\boldsymbol{X}$ is provided by the context; otherwise, if we are not interested in the interpretation of $\boldsymbol{\beta}$, any numerical complement of $\boldsymbol{K}$ will do.*

3.3 Comparison of the two algorithms

The Aitchison-Silvey algorithm requires the inversion of the $(t-1) \times (t-1)$ matrix $\boldsymbol{F}_0$ and the $r \times r$ matrix $(\boldsymbol{H}_0'\boldsymbol{F}_0^{-1}\boldsymbol{H}_0)$; however if we choose, for example, $\boldsymbol{G}$ to be the identity matrix of size t with the first column removed, an explicit inverse exists

$$\boldsymbol{F}^{-1} = \left[n(\operatorname{diag}(\dot{\boldsymbol{\pi}}) - \dot{\boldsymbol{\pi}}\dot{\boldsymbol{\pi}}')\right]^{-1} = n^{-1}\left[\operatorname{diag}(\dot{\boldsymbol{\pi}})^{-1} + \mathbf{1}_{t-1}\mathbf{1}_{t-1}'/(1 - \mathbf{1}_{t-1}'\dot{\boldsymbol{\pi}})\right],$$

where $\dot{\boldsymbol{\pi}}$ denotes the vector $\boldsymbol{\pi}$ with the first element removed.

In the regression algorithm one needs to invert the $(t-1) \times (t-1)$ matrix $\boldsymbol{R}_0^{-1}$ and the $(t-r-1) \times (t-r-1)$ matrix $(\boldsymbol{X}'\bar{\boldsymbol{F}}_0\boldsymbol{X})$. Although the regression algorithm requires two inversions at each step, based upon our implementations it can be slightly faster than the Aitchison-Silvey approach when r is considerably larger than $(t-1)/2$, i.e. there are many more constraints than parameters to be estimated. The particular advantage of the regression algorithm comes in the context of individual-level covariates, as we demonstrate in Section 4.

To give an idea of the time required, we used a random sample from a 4^5 table with 2,000 observations; there were 182 empty cells out of 1,024. Two conditional independence models were fitted, one with 567 and one with 900 constraints. On an average PC, the Aitchison-Silvey algorithm took 26s to run the first case and 39s for the second, while the regression algorithm took 44s and 27s respectively.

3.4 Properties of the algorithms

Detailed conditions for the asymptotic existence of the maximum likelihood estimates of constrained models are given by Aitchison and Silvey (1958); see also Bergsma and Rudas (2002), Theorem 8. Much less is known about existence for finite sample sizes where estimates might fail to exist because of observed zeros. In this case, some elements of $\hat{\boldsymbol{\pi}}$ may converge to $\mathbf{0}$, leading the Jacobian matrix $\boldsymbol{R}$ to become ill-conditioned and making the algorithm unstable.

Concerning the convergence properties of their algorithm, Aitchison and Silvey (1958, p. 827) noted only that it could be seen as a modified Newton algorithm and that similar modifications had been used successfully elsewhere. However, it is clear from the form of the updating equations that if the algorithms converge to some $\boldsymbol{\theta}^*$, then the constraints $\boldsymbol{h}(\boldsymbol{\theta}^*) = \mathbf{0}$ are satisfied, and $\boldsymbol{\theta}^*$ is a stationary point of the constrained likelihood.

To ensure that the stationary point reached by the algorithm is indeed a maximum, one could look at the eigenvalues of the observed information with respect to $\boldsymbol{\beta}$; if these have mixed signs, then we know that the algorithm has converged to a saddle point. An efficient formula for computing the observed information matrix is given in B. Since the log-likelihood of constrained marginal models is not, in general, concave, the algorithm may converge to a local maximum. It might therefore be advisable to apply the algorithm to a range of starting values, in order to check that the maximum is a global one.

3.5 Extension to more general constraints

Occasionally, one may wish to fit general constraints on marginal probabilities without the need to define a marginal log-linear parameterization; an interesting example is provided by the *relational models* of Klimova et al. (2011). They consider constrained models of the form $\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{A}\log(\boldsymbol{M}\boldsymbol{\pi}) = \boldsymbol{0}$, where $\boldsymbol{A}$ is an arbitrary matrix of full row rank. Redefine

$$\boldsymbol{K}' = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}'} = \boldsymbol{A}\,\mathrm{diag}(\boldsymbol{M}\boldsymbol{\pi})^{-1}\boldsymbol{M}\boldsymbol{\Omega}\boldsymbol{G}$$

and note that, because $\boldsymbol{A}$ is not a matrix of row contrasts, $\boldsymbol{h}$ is not homogeneous in $\boldsymbol{\pi}$, and thus the simplifications mentioned in Remark 1 do not apply. If the resulting model is smooth, implying that $\boldsymbol{K}$ is a matrix of full column rank r everywhere in the parameter space, it can be fitted with the ordinary Aitchison-Silvey algorithm. We now show how the same model can also be fitted by a slight extension of the regression algorithm.

Let $\boldsymbol{\theta}_0$ be a starting value and $\bar{\boldsymbol{K}}_0$ be a right inverse of $\boldsymbol{K}'$ at $\boldsymbol{\theta}_0$; consider a first order expansion of the constraints

$$\boldsymbol{h} = \boldsymbol{h}_0 + \boldsymbol{K}_0'(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \boldsymbol{K}_0'(\bar{\boldsymbol{K}}_0\boldsymbol{h}_0 + \boldsymbol{\theta} - \boldsymbol{\theta}_0) = \boldsymbol{0}$$

and let $\boldsymbol{X}_0$ be a matrix that spans the orthogonal complement of $\boldsymbol{K}_0$. Then, with the same order of approximation,

$$\bar{\boldsymbol{K}}_0\boldsymbol{h}_0 + \boldsymbol{\theta} - \boldsymbol{\theta}_0 = \boldsymbol{X}_0\boldsymbol{\beta};$$

by solving the above equation for $\boldsymbol{\theta} - \boldsymbol{\theta}_0$ and substituting into the quadratic approximation of the log-likelihood, we obtain an updating equation similar to (3):

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}_0'\boldsymbol{F}_0\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0'[\boldsymbol{s}_0 + \boldsymbol{F}_0(\bar{\boldsymbol{K}}_0\boldsymbol{h}_0 - \boldsymbol{X}_0\boldsymbol{\beta}_0)].$$

4 Modelling the effect of individual-level covariates

When exogenous covariates are available, it may be of interest to allow the marginal log-linear parameters $\boldsymbol{\eta}$ to depend upon individual-level covariates as in a linear model: $\boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta}$; here the matrix $\boldsymbol{X}_i$ specifies how certain marginal log-linear parameters depend on individual specific information, in addition to structural restrictions such as conditional independence. Let $\boldsymbol{y}_i$, $i = 1, \ldots, n$, be a vector of length t with a 1 in the entry corresponding to the response pattern of the ith individual, and all other values 0; let $\boldsymbol{y}$ be the vector obtained by stacking the vectors $\boldsymbol{y}_i$ one below the other. Alternatively, if the sample size is large and the covariates can take only a limited number of distinct values, $\boldsymbol{y}_i$ may contain the frequency table of the response variables within the sub-sample of subjects with the ith configuration of the covariates; in this case n denotes the number of strata. This arrangement avoids the need to construct a joint contingency table of responses and covariates; in addition the covariate configurations with no observations are simply ignored.

In either case, to implement the Aitchison-Silvey approach, stack the $\boldsymbol{X}_i$ matrices one below the other into the matrix $\boldsymbol{X}$, and let $\boldsymbol{K}$ span the orthogonal complement of $\boldsymbol{X}$; as before, we have to fit the set of constraints $\boldsymbol{K}'\boldsymbol{\eta} = \boldsymbol{0}$. However, while the size of $\boldsymbol{\beta}$ is fixed and, say, equal to q, $\boldsymbol{K}$ will be of size $n(t-1) \times [n(t-1) - q]$, and therefore with n moderately large, the problem is almost infeasible.

In the regression formulation the complexity of the problem is at worst linear in n, as we now show. Let $\boldsymbol{\theta}_i$ denote the vector of canonical parameters for the ith individual and $l(\boldsymbol{\theta}_i) = \boldsymbol{y}'\boldsymbol{G}\boldsymbol{\theta}_i - \log[\boldsymbol{1}_t'\exp(\boldsymbol{G}\boldsymbol{\theta}_i)]$ be the contribution to the log-likelihood. Note that $\boldsymbol{X}_i$ need not be of full column rank, a property which must instead hold for the matrix $\boldsymbol{X}$; for this reason our assumptions are much weaker than those used by Lang (1996), and allow for more flexible models.

Both the quadratic and the linear approximations must be applied at the individual level; thus we set $\boldsymbol{\theta}_i - \boldsymbol{\theta}_{i0} = \boldsymbol{R}_{i0}(\boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{\eta}_{i0})$, and the log-likelihood becomes

$$\sum_{i=1}^{n} l(\boldsymbol{\theta}_i) \cong -\frac{1}{2}\sum_{i=1}^{n}[\boldsymbol{R}_{i0}(\boldsymbol{X}_i\boldsymbol{\delta} - \boldsymbol{\gamma}_{i0}) - \boldsymbol{F}_{i0}^{-1}\boldsymbol{s}_{i0}]'\boldsymbol{F}_{i0}[\boldsymbol{R}_{i0}(\boldsymbol{X}_i\boldsymbol{\delta} - \boldsymbol{\gamma}_{i0}) - \boldsymbol{F}_{i0}^{-1}\boldsymbol{s}_{i0}],$$

where $\boldsymbol{\gamma}_{i0} = \boldsymbol{\eta}_{i0} - \boldsymbol{X}_i\boldsymbol{\beta}_0$, $\boldsymbol{s}_i = \boldsymbol{G}'(\boldsymbol{y}_i - \boldsymbol{\pi}_i)$ and $\boldsymbol{F}_i = \boldsymbol{G}'\boldsymbol{\Omega}_i\boldsymbol{G}$.

Direct calculations lead to the updating expression

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = \left(\sum_i \boldsymbol{X}_i'\boldsymbol{W}_i\boldsymbol{X}_i\right)^{-1}\left[\sum \boldsymbol{X}_i'(\boldsymbol{W}_i\boldsymbol{\gamma}_{i0} + \boldsymbol{R}_{i0}'\boldsymbol{s}_{i0})\right],$$

where $\boldsymbol{W}_i = \boldsymbol{R}_{i0}'\boldsymbol{G}'\boldsymbol{\Omega}_{i0}\boldsymbol{G}\boldsymbol{R}_{i0}$.

The complexity of this procedure is therefore at worst linear in the number of observations, and potentially much less if the covariates are discrete.

For an application of the method described above to social mobility tables see Dardanoni et al. (2012). Social mobility tables are cross classifications of subjects according to their social class (columns) and that of their fathers (rows). The hypothesis of equality of opportunity would imply that the social class of sons is independent of that of their fathers. Mediating covariates may induce positive dependence between the social classes of fathers and sons, leading to the appearance of limited social mobility; to assess this, Dardanoni et al. (2012) fitted a model where the vector of marginal parameters for each son-father pair was allowed to depend on individual covariates, including the father's age, the results of cognitive and non-cognitive test scores taken by the son at school, and his academic qualifications. The analysis, based on the UK's National Child Development Survey, indicated that positive association was present even after controlling for a rich set of covariates, thus demonstrating inequality of opportunity for the sons based on their fathers' occupations.

5 L_1-penalized parameters

Evans (2011) shows that, in the context of marginal log-linear parameters, consistent model selection can be performed using the so-called adaptive lasso. Since the adaptive lasso uses L_1-penalties, we might therefore be interested in relaxing the equality constraints discussed above to a penalization framework, in which we maximize the penalized log-likelihood

$$\phi(\boldsymbol{\theta}) \equiv l(\boldsymbol{\theta}) - \sum_{j=1}^{t-1}\nu_j|\eta_j(\boldsymbol{\theta})|,$$

for some vector of penalties $\boldsymbol{\nu} = (\nu_j) \geq \mathbf{0}$.

The advantage of penalties of this form is that one can obtain parameter estimates which are exactly zero (Tibshirani, 1996), and therefore perform model selection without the need to fit many models separately. For now, assume that no equality constraints hold for $\boldsymbol{\eta}$, so we can take $\boldsymbol{X}$ to be the identity, and $\boldsymbol{\beta} = \boldsymbol{\eta}$. This gives the quadratic form

$$Q(\boldsymbol{\eta}) = -\frac{1}{2}[\boldsymbol{R}_0(\boldsymbol{\eta} - \boldsymbol{\eta}_0) - \boldsymbol{F}_0^{-1}\boldsymbol{s}_0]'\boldsymbol{F}_0[\boldsymbol{R}_0(\boldsymbol{\eta} - \boldsymbol{\eta}_0) - \boldsymbol{F}_0^{-1}\boldsymbol{s}_0]$$

approximating $l(\boldsymbol{\theta})$ as before. Then ϕ is approximated by

$$\tilde{\phi}(\boldsymbol{\eta}) \equiv -\frac{1}{2}[\boldsymbol{R}_0(\boldsymbol{\eta} - \boldsymbol{\eta}_0) - \boldsymbol{F}_0^{-1}\boldsymbol{s}_0]'\boldsymbol{F}_0[\boldsymbol{R}_0(\boldsymbol{\eta} - \boldsymbol{\eta}_0) - \boldsymbol{F}_0^{-1}\boldsymbol{s}_0] - \sum_j \nu_j|\eta_j|,$$

and we can attempt to maximize ϕ by repeatedly solving the sub-problem of maximizing $\tilde{\phi}$. Now, because the quadratic form $Q(\boldsymbol{\eta})$ is concave and differentiable, and the absolute value function $|\cdot|$ is concave, coordinate-wise ascent is guaranteed to find a local maximum of $\tilde{\phi}$ (Tseng, 2001). Coordinate-wise ascent cycles through $j = 1, 2, \ldots, t-1$, at each step minimizing

$$-\frac{1}{2}[\boldsymbol{R}_0(\boldsymbol{\eta} - \boldsymbol{\eta}_0) - \boldsymbol{F}_0^{-1}\boldsymbol{s}_0]'\boldsymbol{F}_0[\boldsymbol{R}_0(\boldsymbol{\eta} - \boldsymbol{\eta}_0) - \boldsymbol{F}_0^{-1}\boldsymbol{s}_0] - \nu_j|\eta_j|$$

with respect to η_j, with $\eta_1, \ldots, \eta_{j-1}, \eta_{j+1}, \ldots, \eta_{t-1}$ held fixed. This is solved just by taking

$$\eta_j = \operatorname{sign}(\check{\eta})(|\check{\eta}| - \nu_j)_+,$$

where $a_+ = \max\{a, 0\}$, and $\check{\eta}_j$ minimizes Q with respect to η_j (Friedman et al., 2010). This approach to the sub-problem may require a large number of iterations, but it is extremely fast in practice because each step is so simple. If the overall algorithm converges, then by a similar argument to that of Section 3.4, together with the fact that $\tilde{\phi}$ has the same supergradient as ϕ at $\boldsymbol{\eta} = \boldsymbol{\eta}_0$, we see that we must have reached a local maximum of ϕ.

Since penalty selection for the lasso and adaptive lasso is typically performed using computationally intensive procedures such as cross validation, its implementation makes fast algorithms such as the one outlined above essential.

A Proof of Equivalence of the Algorithms

Lemma 1. *For matrices $\boldsymbol{X}$ and $\boldsymbol{K}$, let the columns of $\boldsymbol{X}$ span the orthogonal complement of the space spanned by the columns of $\boldsymbol{K}$. Then for any symmetric and positive definite matrix $\boldsymbol{W}$*

$$\boldsymbol{W}^{-1} - \boldsymbol{W}^{-1}\boldsymbol{K}(\boldsymbol{K}'\boldsymbol{W}^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}'\boldsymbol{W}^{-1} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'. \tag{6}$$

Proof. Let $\boldsymbol{U} = \boldsymbol{W}^{-1/2}\boldsymbol{K}$ and $\boldsymbol{V} = \boldsymbol{W}^{1/2}\boldsymbol{X}$ and note that $\boldsymbol{U}'\boldsymbol{V} = \boldsymbol{K}'\boldsymbol{X} = \boldsymbol{0}$, then (6) follows from the identity $\boldsymbol{U}(\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}' + \boldsymbol{V}(\boldsymbol{V}'\boldsymbol{V})^{-1}\boldsymbol{V}' = \boldsymbol{I}$. □

Proof of Proposition 1. Recall $\bar{\boldsymbol{s}} = \boldsymbol{R}'\boldsymbol{s}$ and $\bar{\boldsymbol{F}} = \boldsymbol{R}'\boldsymbol{F}\boldsymbol{R}$, and note that

$$\boldsymbol{H}'\boldsymbol{F}^{-1}\boldsymbol{H} = \boldsymbol{K}'\boldsymbol{R}^{-1}\boldsymbol{F}^{-1}(\boldsymbol{R}^{-1})'\boldsymbol{K} = \boldsymbol{K}'\bar{\boldsymbol{F}}^{-1}\boldsymbol{K};$$

using this in the updating equation (2) enables us to rewrite it as

$$\begin{aligned}\boldsymbol{R}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = [&\bar{\boldsymbol{F}}_0^{-1} - \bar{\boldsymbol{F}}_0^{-1}\boldsymbol{K}(\boldsymbol{K}'\bar{\boldsymbol{F}}_0^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}'\bar{\boldsymbol{F}}_0^{-1}]\bar{\boldsymbol{s}}_0 + \\ &- \bar{\boldsymbol{F}}_0^{-1}\boldsymbol{K}(\boldsymbol{K}'\bar{\boldsymbol{F}}_0^{-1}\boldsymbol{K})\boldsymbol{K}\bar{\boldsymbol{F}}_0^{-1}\bar{\boldsymbol{F}}_0\boldsymbol{\eta}_0.\end{aligned} \tag{7}$$

Set $\boldsymbol{W} = \bar{\boldsymbol{F}}_0$ and note that (6) may be substituted into the first component of (7) and that its equivalent formulation

$$\bar{\boldsymbol{F}}_0^{-1}\boldsymbol{K}(\boldsymbol{K}'\bar{\boldsymbol{F}}_0^{-1}\boldsymbol{K})^{-1}\boldsymbol{K}'\bar{\boldsymbol{F}}_0^{-1} = \bar{\boldsymbol{F}}_0^{-1} - \boldsymbol{X}(\boldsymbol{X}'\bar{\boldsymbol{F}}_0\boldsymbol{X})^{-1}\boldsymbol{X}'$$

may be substituted into the second component, giving

$$\boldsymbol{R}_0^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) = \boldsymbol{X}(\boldsymbol{X}'\bar{\boldsymbol{F}}_0\boldsymbol{X})^{-1}\boldsymbol{X}'\bar{\boldsymbol{s}}_0 - \boldsymbol{\eta}_0 + \boldsymbol{X}(\boldsymbol{X}'\bar{\boldsymbol{F}}_0\boldsymbol{X})^{-1}\boldsymbol{X}'\bar{\boldsymbol{F}}_0\boldsymbol{\eta}_0.$$

This is easily seen to be the same as combining equations (3) and (4). □

B Computation of the observed information matrix

Lemma 2. *Suppose that $\boldsymbol{A}$ is a $p \times q$ matrix, and that $\boldsymbol{y}$, $\boldsymbol{b}$, $\boldsymbol{x}$ and $\boldsymbol{u}$ are column vectors with respective lengths q, p, k and r. Then if $\boldsymbol{A}$ and $\boldsymbol{b}$ are constant,*

$$\frac{\partial}{\partial \boldsymbol{x}'} \operatorname{diag}(\boldsymbol{A}\boldsymbol{y})\boldsymbol{b} = \operatorname{diag}(\boldsymbol{b})\boldsymbol{A}\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{u}'}\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}'}. \tag{8}$$

Proof.

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{x}'} \operatorname{diag}(\boldsymbol{A}\boldsymbol{y})\boldsymbol{b} &= \frac{\partial}{\partial \boldsymbol{u}'} \operatorname{diag}(\boldsymbol{A}\boldsymbol{y})\boldsymbol{b}\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}'} \\ &= \left(\operatorname{diag}(\boldsymbol{A}\boldsymbol{y}_{u1})\boldsymbol{b}, \ \cdots \ \operatorname{diag}(\boldsymbol{A}\boldsymbol{y}_{uh})\boldsymbol{b}\right)\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}'} \\ &= \left(\operatorname{diag}(\boldsymbol{b})\boldsymbol{A}\boldsymbol{y}_{u1}, \ \cdots \ \operatorname{diag}(\boldsymbol{b})\boldsymbol{A}\boldsymbol{y}_{uh}\right)\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}'} \\ &= \operatorname{diag}(\boldsymbol{b})\boldsymbol{A}\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{u}'}\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}'}.\end{aligned}$$

□

The observed information matrix is minus the second derivative of the log-likelihood with respect to $\boldsymbol{\beta}$, that is

$$\begin{aligned}-\frac{\partial}{\partial \beta'}\left[\frac{\partial l(\boldsymbol{\theta})}{\partial \beta}\right] &= -\frac{\partial}{\partial \beta'}\boldsymbol{X}'\boldsymbol{R}'\boldsymbol{G}'(\boldsymbol{y} - n\boldsymbol{\pi}) = -\left[\frac{\partial}{\partial \boldsymbol{\theta}'}\boldsymbol{X}'\boldsymbol{R}'\boldsymbol{G}'(\boldsymbol{y} - n\boldsymbol{\pi})\right]\boldsymbol{R}\boldsymbol{X} \\ &= n\boldsymbol{X}'\boldsymbol{R}'\boldsymbol{G}'\boldsymbol{\Omega}\boldsymbol{G}\boldsymbol{R}\boldsymbol{X} - \boldsymbol{X}'\frac{\partial \boldsymbol{R}'}{\partial \boldsymbol{\theta}'}\boldsymbol{G}'(\boldsymbol{y} - n\boldsymbol{\pi})\boldsymbol{R}\boldsymbol{X}.\end{aligned}$$

Since $\boldsymbol{s}$ depends on $\boldsymbol{\theta}$ through both $(\boldsymbol{y} - n\boldsymbol{\pi})$ and $\boldsymbol{R}$, the above derivative has two main components, where the one obtained by differentiating $(\boldsymbol{y} - n\boldsymbol{\pi})$ is minus the expected information. Using the well known expression for the derivative of an inverse matrix, it only remains to compute

$$\boldsymbol{X}'\frac{\partial \boldsymbol{R}'}{\partial \boldsymbol{\theta}'}\boldsymbol{G}'(\boldsymbol{y} - n\boldsymbol{\pi})\boldsymbol{R}\boldsymbol{X} = \boldsymbol{X}'\boldsymbol{R}'\frac{\partial \boldsymbol{R}'^{-1}}{\partial \boldsymbol{\theta}'}\boldsymbol{R}'\boldsymbol{G}'(\boldsymbol{y} - n\boldsymbol{\pi})\boldsymbol{R}\boldsymbol{X} = \boldsymbol{A}\frac{\partial \boldsymbol{R}'^{-1}}{\partial \boldsymbol{\theta}'}\boldsymbol{b}\boldsymbol{R}\boldsymbol{X}$$

where $\boldsymbol{A} = \boldsymbol{X}'\boldsymbol{R}'$ and $\boldsymbol{b} = \boldsymbol{R}'\boldsymbol{G}'(\boldsymbol{y} - n\boldsymbol{\pi})$, giving

$$= \boldsymbol{A}\boldsymbol{G}'\frac{\partial[\operatorname{diag}(\boldsymbol{\pi})\boldsymbol{M}'\operatorname{diag}(\boldsymbol{M}\boldsymbol{\pi})^{-1}]}{\partial \boldsymbol{\theta}'}\boldsymbol{C}'\boldsymbol{b}\boldsymbol{R}\boldsymbol{X}.$$

By two applications of (8), this is

$$\begin{aligned}&\boldsymbol{A}\boldsymbol{G}'\big[\operatorname{diag}(\boldsymbol{M}'\operatorname{diag}(\boldsymbol{M}\boldsymbol{\pi})^{-1}\boldsymbol{C}'\boldsymbol{b}) \\ &\qquad - \operatorname{diag}(\boldsymbol{\pi})\boldsymbol{M}'\operatorname{diag}(\boldsymbol{C}'\boldsymbol{b})\operatorname{diag}(\boldsymbol{M}\boldsymbol{\pi})^{-2}\boldsymbol{M}\big]\boldsymbol{\Omega}\boldsymbol{G}\boldsymbol{R}\boldsymbol{X}.\end{aligned}$$

References

A. Agresti. *Categorical data analysis.* John Wiley and Sons, 2002.

J. Aitchison and S. D. Silvey. Maximum-likelihood estimation of parameters subject to restraints. *Ann. Math. Stat.*, 29(3):813–828, 1958.

F. Bartolucci, R. Colombi, and A. Forcina. An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints. *Statist. Sinica*, 17(2):691, 2007.

W. Bergsma, M. Croon, and J. A. Hagenaars. *Marginal Models: For Dependent, Clustered, and Longitudinal Categorial Data.* Springer Verlag, 2009.

W. P. Bergsma. *Marginal models for categorical data.* Tilburg University Press, Tilburg, 1997.

W. P. Bergsma and T. Rudas. Marginal models for categorical data. *Ann. Statist.*, 30(1): 140–159, 2002.

R. Colombi and A. Forcina. Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, 88(4):1001–1019, 2001.

V. Dardanoni, M. Fiorini, and A. Forcina. Stochastic monotonicity in intergenerational mobility tables. *J. Appl. Economtrics*, 27:85–107, 2012.

R. J. Evans. *Parametrizations of discrete graphical models.* PhD thesis, University of Washington, 2011.

R. J. Evans and T. S. Richardson. Marginal log-linear parameters for graphical markov models. arXiv:1105.6075, 2011.

A. Forcina, M. Lupparelli, and G. M. Marchetti. Marginal parameterizations of discrete models defined by a set of conditional independencies. *Journ. Mult. Analysis*, 101(10): 2519–2527, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Soft.*, 33(1), 2010.

G. F. V. Glonek and P. McCullagh. Multivariate logistic models. *J. R. Statist. Soc. B*, 57 (3):533–546, 1995.

A. Klimova, T. Rudas, and A. Dobra. Relational models for contingency tables. arXiv:1102.5390, 2011.

J.B. Lang. Maximum likelihood methods for a generalized class of log-linear models. *Ann. Statist.*, 24(2):726–752, 1996.

K. Y. Liang, S. L. Zeger, and B. Qaqish. Multivariate regression analyses for categorical data. *J. R. Statist. Soc. B*, 54(1):3–40, 1992.

P. McCullagh and J. A. Nelder. *Generalized linear models.* Chapman & Hall/CRC, 1989.

T. Rudas, W. P. Bergsma, and R. Németh. Marginal log-linear parameterization of conditional independence models. *Biometrika*, 97(4):1006–1012, 2010.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. B*, 58(1):267–288, 1996.

P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optim. Theory Appl.*, 109(3):475–494, 2001.